Genomics Proteomics Bioinformatics 11 (2013) 61–65

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

ORIGINAL RESEARCH

Shigella Strains Are Not Clones of *Escherichia coli* but Sister Species in the Genus *Escherichia*

Guanghong Zuo [†], **Zhao Xu** [§], **Bailin Hao** [*]

T-Life Research Center and Department of Physics, Fudan University, Shanghai 200433, China

Received 28 October 2012; accepted 5 November 2012
Available online 29 December 2012

KEYWORDS

Shigella;
Escherichia coli;
Prokaryote phylogeny and taxonomy;
Composition vector;
CVTree

Abstract *Shigella* species and *Escherichia coli* are closely related organisms. Early phenotyping experiments and several recent molecular studies put *Shigella* within the species *E. coli*. However, the whole-genome-based, alignment-free and parameter-free CVTree approach shows convincingly that four established *Shigella* species, *Shigella boydii*, *Shigella sonnei*, *Shigella felxneri* and *Shigella dysenteriae*, are distinct from *E. coli* strains, and form sister species to *E. coli* within the genus *Escherichia*. In view of the overall success and high resolution power of the CVTree approach, this result should be taken seriously. We hope that the present report may promote further in-depth study of the *Shigella-E. coli* relationship.

Introduction

Although description of bacillary dysentery can be traced back in ancient records, the aetiologic agent was recognized only in late 19th century. In 1898 Shiga gave a detailed description of what was called *Bacillus dysenteriae*, which was assigned a new genus *Shigella* later on. Four *Shigella* species, *Shigella dysenteriae*, *Shigella boydii*, *Shigella sonnei* and *Shigella flexineri*, have been identified and listed in several editions of the Bergey's Manual, including the latest one [1]. However, it has been known since the 1970s that DNA–DNA reassociation studies and a few other phenotyping experiments could not distinguish these species from *Escherichia coli* strains (see, *e.g.*, [2,3]). Therefore, these *Shigella* organisms and *E. coli* were considered "one species genetically" [4].

Recent molecular studies further validated the closeness of the *Shigella* species and *E. coli*. Pupo et al. referred to all *Shigella* strains as "forms of *E. coli*" by using multilocus enzyme electrophoresis (MLEE) and a housekeeping gene sequence study [5]. Later on these authors simply called the *Shigella* species "clones of *E. coli*" [6], suggesting that the *Shigella* species may have originated from different ancestral strains of *E. coli* and have undergone convergent evolution to their present status. Ogura et al. [7] further constructed a neighbor-joining tree by using concatenated nucleotide sequences of 345 orthologous CDS groups from 25 sequenced strains (19 *E. coli* and 6 *Shigella*). The *Shigella* strains again were assigned as *E. coli* strains [7].

As sequences of more and more complete genomes become available, the use of housekeeping genes has been extended to "core genome". For example, 2034 genes from the "core genome" were selected to construct phylogenetic relationships (22

* Corresponding author.
E-mail: hao@mail.itp.ac.cn (Hao B).
† Present address: Shanghai Institute of Applied Physics, Chinese Academy of Sciences, Shanghai 201800, China.
§ Present address: Applied Biosystems, Beijing 100027, China.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

ELSEVIER Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.gpb.2012.11.002

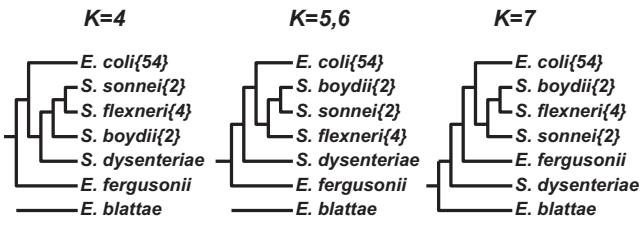

Figure 1 The *Escherichia-Shigella* branch in CVTrees at $K = 4$–7, respectively

Numerals in parentheses indicate the number of genomes in a branch.

E. coli and 7 *Shigella* in [8], see their Figure 3; or 49 *E. coli* and 7 *Shigella* in [9], see their Figure 1). In all the aforementioned studies, the *Shigella* species were mixed up with the *E. coli* strains. Investigation using 16S rRNA segments and *in silico* multilocus sequence typing (MLST) based on a small number of housekeeping genes [10] led to more scattered results. Even a recent "alignment-free" study using so-called feature frequency profiles [11] placed the *Shigella* species into the *E. coli* strains. There has been a consensus that the *Shigella* species are indeed *E. coli* strains and the nomenclature of the genus *Shigella* and species included within this genus has been kept for historical and medical reasons. No wonder that the *Shigella* strains were called *E. coli* "in disguise" [12] or "Machiavellian masqueraders" [13].

On the other hand, it is curious enough that despite the genetic closeness of the *Shigella* species and *E. coli* strains, certain distinctive "morphological" features do show up. Besides the diagnosable clinical difference of the dysentery they cause, there are some other observable dissimilarities. For example, *E. coli* strains usually have flagella and are motile, but *Shigella* species do not, though their flagella genes may express under some rare, yet not fully-understood circumstances [14].

As any phylogenetic conclusion drawn from the analysis of a selected set of sequence segments or genes cannot be unambiguously convincing, there is an urgent need for methods that are not based on any special choice of sequences or genes and that do not require any adjustment of parameters. A few years ago we developed such a whole-genome-based, alignment-free, and parameter-free method [15,16], called CVTree in accordance with the name of the public domain web server CVTree [17,18]. The CVTree results clearly show that the four *Shigella* species as well as all the *E. coli* strains are well-defined monophyletic clusters of their own; the *Shigella* species are not clones of *E. coli*, but members of the genus *Escherichia* on the same footing as the *E. coli* species. The only possible change in nomenclature concerns merging the two genera, *Shigella* and *Escherichia*, into one genus, but not absorbing the *Shigella* strains into the *E. coli* species.

Though challenging to the current consensus described above, in view of the overall success of the CVtree approach and its high resolution power (see, *e.g.*, [19,20]), this conclusion cannot be simply ignored or negated.

Results and discussion

We shall not reproduce the 2070-population CVTrees in this report. An interested reader may generate the result by going to the CVTree web server and ticking the appropriate names in the list of built-in genomes. We base our discussion on collapsed subtrees cut from the 2070-genome CVTrees. **Figure 1** shows the *Escherichia-Shigella* branch in CVTrees at different Ks in the "collapsed-tree" notation. At $K = 3$ (not shown), there was a monophyletic *Shigella*{9} branch, but one of the *E. coli* genome (one of the "engineered" Waksman strain KO11LF) escaped from the *Escherichia* cluster, violating the monophyleticity of the latter. The situation improves for $K > 3$. Figure 1 and **Figure 2** provide examples of convergence of the branching scheme with increasing K. $K = 4$ is better than $K = 3$ and $K = 5$ and 6 are the best, while $K = 7$ may be slightly worse (see our previous publications [19,20]). An important and consistent fact consists in that all the *Shigella* species as well as all the *E. coli* strains form monophyletic clusters of their own. The *Shigella* species are never included in the *E. coli* branch. *Shigella* species are sister species to *E. coli* but not strains within the *E. coli* monophyletic branch. We note that the position of the newly-sequenced genome of *E. blattae* in Figure 1 requires further study, but this does not affect the *E. coli-Shigella* relationship, which is the main concern of this work.

The results of this whole-genome-based and alignment-free CVTree analysis convincingly reconcile the seeming contradiction between the genetic closeness and the "morphological" differences mentioned in the "Introduction" section.

The grouping of the 54 *E. coli* strains within the monophyletic cluster (Figure 2) reflects the evolution and taxonomy of the strains in much the same way as revealed in many previous studies using different methods (see, *e.g.*, [7–11]). It is remarkable that the six monophyletic clusters within the *E. coli*{54} branch agree well with the phylogroups commonly used to characterize the *E. coli* population. This is why we use the phylogroup labels A, B1 (split into B1a and B1b), B2, D, and E to name the six groups in Table S1. Group A contains the commensal strains and their derivatives: the K-12 strains (MG1655, W3110, BW2952, DH1 and DH10B) and the B strains (BL21 and REL606) [21]. The Waksman strains (W [22] and its derivative KO11FL [23]) and the commensal strains IAI1 [24], SE11 [25], enterotoxigenic (ETEC) E24377A and enteroaggregative (EAEC) 55989 form group B1b [26]. The virulent enterohemorrhagic *E. coli* (EHEC) O157:H7 strains [27–30] and their O55:H7 precursors [8,31] form a monophyletic cluster E. The three non-O157 EHEC

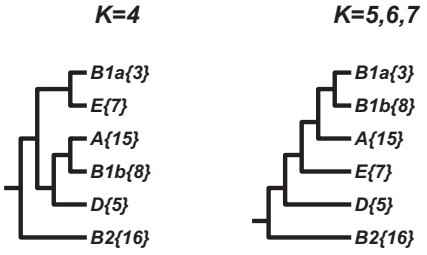

Figure 2 The monophyletic *E. coli*{54} branch consists of six subclusters

These six monophyletic clusters agree well with the phylogroups commonly used to characterize the *E. coli* population and are therefore labeled accordingly as A, B1a, B1b, B2, D, and E. Numerals in parentheses give the number of genomes in each group as indicated in the first column of Table S1.

phylogroup B1 strains (O26, O103 and O111) [7] join the other phylogroup B1a. The many uropathogenic (UPEC) strains of phylogroup B2 form a large lowermost cluster B2. Note that though the separation of *E. coli* strains into clusters agrees basically with [7–11] and other studies, the *Shigella* strains always stay clearly outside the *E. coli* monophyletic branch. As the main aim of this report is to emphasize the fact that *Shigella* species are members of the genus *Escherichia*, not strains of *E. coli*, we postpone the detailed comparison of the inner structure of the subclusters within the *E. coli*{54} branch to a later publication.

We mention in passing that a similar story is told by the *Yersinia pestis* and *Y. pseudotuberculosis* strains in the CVTrees. Strains from these two species could not be distinguished by DNA-DNA hybridization. Therefore, a proposal was made to combine these two species into one. However, "... the change was rejected by the Judicial Commission because of possible danger to public health if there was confusion regarding *Y. pestis*, the plague bacillus" [32]. In the same 2070-population CVTrees, we see *Yersinia*{19} K3K4K5K6K7, *Y. pestis*{12} K3K4K5K6K7, *Y. pseudotuberculosis*{4} K4K5K6K7 and *Y. enterocolitica*{3} K3K4K5K6K7. Consequently, the genus *Yersinia* and the three species therein are all well-defined and there is no worry for the taxonomic Judicial Commission.

It should be pointed out that we did not carry out any case study for a group of selected organisms. Instead, we generated CVTrees for all 2062 *Archaea* and *Bacteria* genomes, cut and scrutinized the interested branch. The results demonstrated the high resolution power of CVTrees at the subspecies level and below. This resolution is beyond the reach of the 16S rRNA analysis. Concatenation of a large number of nucleotide or protein sequences such as done in [5–9] may lead to seemingly comparable resolution, but the somewhat subjective selection of sequences or genes brings about ambiguity and makes the conclusion less convincing.

With the progress of the new generations of sequencing techniques, the cost of sequencing a bacterial genome will soon drop below that of an average phenotyping experiment and the number of sequenced prokaryotic genomes keeps growing rapidly. Among the genomes released at the NCBI FTP site (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/), there are more and more strains coming from the same species. For example, for the time being, complete genome sequences of ten or more strains are available for *Chlamydia trachomatis*, *Corynebacterium diphtheriae*, *Helicobacter pylori*, *Salmonella enterica*, *Staphylococcus aureus*, *Streptococcus pneumoniae*, *Streptococcus pyogenes*, *Sulfolobus islandicus*, *Y. pestis*, *etc*. Once the genomes have been sequenced, there is no additional cost to getting the interrelationship of the strains by simply submitting the genomes to the CVTree Web Server. We encourage researchers to try out this convenient and effective tool.

Materials and methods

Since the CVTree approach has been described repeatedly in previous publications (see [15–20] and references therein), we only give a brief summary in order to introduce notations and concepts needed in what follows.

CVTree is a whole-genome-based approach. It makes use of all the protein products encoded in a given genome. In this way it circumvents the problem of lateral gene transfer (LGT) as LGT and lineage-dependent gene loss are merely mechanisms of genome evolution. User avoids the tedious task of finding orthologous proteins as well, since all genomes are orthologous as they are descended from a common ancestor.

The methodology of CVTree must be alignment-free due to the extreme diversity of bacterial genomes in their size and gene content. By using a sliding window of width K, a primary protein sequence made of L amino acids is replaced by $(L - K + 1)$ peptides of length K. The number of K-peptides from all the protein products in a genome is counted and these counts are put in lexicographic order of all possible K-peptides over the 20 amino acid letters to form a raw composition vector (CV) of dimension 20^K. Then a random background caused by neutral mutations is subtracted from each raw count to highlight the role of natural selection by using a $(K - 2)$th order Markovian prediction formula. The subtraction procedure is crucial to the success of CVTree. A recalculated CV represents a species and a dissimilarity/distance measure is defined between each pair of CVs. Then a phylogenetic tree is constructed by using the standard neighbor-joining algorithm which has been proved to be a robust quartet-based method [33].

Being alignment-free renders the method parameter-free, as sequence alignment involves many parameters embodied in the elements of scoring matrices and gap penalties. The peptide length K is not a parameter. Longer Ks make emphasis on species-specificity, while shorter Ks reflect common features between different species. We never adjust K value. Five trees are calculated for $K = 3$–7 (there is no need to go beyond $K = 7$) and the improved agreement of the tree topology with taxonomy when K increases provides an additional angle to evaluate the quality of the resulted trees.

In order to facilitate the use of the new method by biologists practitioners, a web server entitled CVTree was published in 2004 [17]. A significantly-improved version was released in 2009 [18]. Just by entering the URL (http://tlife.fudan.edu.cn/cvtree/) into a browser, user can enjoy playing with CVTree. The built-in dataset is updated automatically in the beginning of each month from the NCBI FTP site. Users may also upload their own data to CVTree, these data will be kept only for 48 h after the last run of the job. The results may be displayed online or sent back to users by email. In the latter case, there is a directory named Collapsed-trees with many files in Newick (.nwk) or plain text format. The notion of collapsed trees requires special explanation.

Although statistical re-sampling methods such as bootstrap or jackknife have been designed to check the stability and self-consistency of the CVTree results [34], the CVTrees are verified by direct comparison with prokaryote systematics at all taxonomic ranks from domain down to genera and species. In doing so, the monophyleticity of a branch is taken as a guideline. When all genomes from one and the same taxon in the input dataset appear in the same branch and no other genomes fall in, one may collapse the branch to a single leave named after the taxon. For example, *Escherichia_coli*{54} means that all 54 *E. coli* genomes appear in a monophyletic branch at a given K. In fact, we have the *E. coli* strains making a monophyletic branch at all K-values from 4 to 7, which is denoted as "*Escherichia_coli*{54} K4K5K6K7". For the time being, this kind of "convergence lists" has to be obtained by manual inspection of the corresponding files returned via email by the CVTree web server. Automatic generation of such lists at all

taxonomic ranks will be implemented in the next release of the CVTree web server.

Throughout this paper we use the abbreviation CVTree to denote the method, the CVTree web server, and the phylogenetic tree obtained by using the CVTree web server.

In the present study, we have used all the prokaryote genomes released at the NCBI FTP site as of 30 September 2012, excluding 14 tiny highly-degenerated genomes of bacterial endosymbiont bacteria. The 54 *E. coli* genomes, listed in Table S1 in the Supplementary material, are divided into six groups, corresponding to the six monophyletic clusters within the monophyletic *E. coli*{54} branch in CVTrees for $K = 4$–7 (see Figure 2). We note that 49 [9] and 53 [10] *E. coli* genomes from GenBank were used, respectively. There are minor differences in the lists as we used all the genomes released by NCBI with accession number starting with NC_ in order to have better comparability. The 9 *Shigella* genomes used in the present study are listed in Table S2.

When constructing the phylogenetic trees, we used all 133 Archaea genomes and 1929 Bacteria genomes, including the 54 *E. coli* and 9 *Shigella* genomes. We excluded 14 tiny highly-degenerated genomes of endosymbiont bacteria (*Candidatus* Carsonella, *C.* Hodgkinia, *C.* Sulcia, *C.* Tremblaya, and *C.* Zinderia), as they would violate the trifurcation of the three main domains of life. Eight Eukarya genomes were included as outgroups. Altogether it led to a treeing job with $133 + 1929 + 8 = 2070$ population.

Authors' contributions

GZ and BH posed the problem. ZX maintained the CVTree web server. GZ, ZX and BH collected data and performed the calculation. GZ and BH analyzed the results and wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors have declared that no competing interests exist.

Acknowledgements

This work was supported by the National Basic Research Program of China (973 Project, Grant No. 2007CB814800 and 2013CB834100), the Shanghai Leading Academic Discipline Project (Grant No. B111), the National Key Laboratory of Applied Surface Physics and the Department of Physics, Fudan University. BH thanks Dr. Kui Lin for a discussion on *E. coli* phylogroups.

Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.gpb.2012.11.002.

References

[1] Bergey's Manual Trust. Bergey's Manual of Systematic Bacteriology. 2nd ed., vols. 1–5. Berlin, Heidelberg, New York: Springer-Verlag; 2001–2012.

[2] Brenner DJ, Fanning GR, Skerman FJ, Falkow S. Polynucleotide sequence divergence among strains of *Escherichia coli* and closely related organisms. J Bacteriol 1972;109:953–65.

[3] Brenner DJ, Fanning GR, Milkos GV, Steigerwalt AG. Polynucleotide sequence relatedness among *Shigella* species. Int J Syst Bacteriol 1973;23:1–7.

[4] Brenner DJ. Introduction to the family *Enterobacteriaceae*, Chapter 88. In: Starr MP, Stolp H, Truper HG, Balows A, Schleger HC, editors. The prokaryotes. Berlin, Heidelberg, New York: Springer-Verlag; 1981. p. 1105–27.

[5] Pupo GM, Karaolis DKR, Lan R, Reeves PR. Evolutionary relationships among pathogenic and non-pathogenic *Escherichia coli* strains inferred from multilocus enzyme electrophoresis and mdh sequence studies. Infect Immun 1997;65:2685–92.

[6] Pupo GM, Lan R, Reeves PR. Multiple independent origin of Shigella clones of *Escherichia coli* and convergent evolution of many of their characteristics. Proc Natl Acad Sci U S A 2000;97:10567–72.

[7] Ogura Y, Ooka T, Iguchi Y, Toh A. Comparative genomics reveal the mechanism of the parallel evolution of O157 and non-O157 enterohemorrhagic *Escherichia coli*. Proc Natl Acad Sci U S A 2009;106:17939–44.

[8] Zhou Z, Li X, Liu B, Beutin L, Xu J, Ren Y, et al. Derivation of *Escherichia coli* O157:H7 from its O55:H7 precursor. PLoS One 2010;5:e8700.

[9] Reeves PR, Liu B, Zhou Z, Li D, Guo D, Ren Y, et al. Rates of mutation and host transmission for an *Escherichia coli* clone over 3 years. PLoS One 2011;6:e26907.

[10] Lukjancenko O, Wassenaar TM, Ussery DW. Comparison of 61 sequenced *Escherichia coli* genomes. Microb Ecol 2010;60:708–20.

[11] Sims GE, Kim SH. Whole-genome phylogeny of the *Escherichia coli*/*Shigella* group by feature frequency profiles (FFPs). Proc Natl Acad Sci U S A 2011;108:8329–34.

[12] Lan R, Reeves PR. *Escherichia coli* in disguise: molecular origin of *Shigella*. Microbes Infect 2002;4:1125–32.

[13] Johnson JR. *Shigella* and *Escherichia coli* at the crossroads: machiavellian masqueraders or taxonomic treachery? J Med Microbiol 2000;49:583–5.

[14] Giron JA. Expression of flagella and motility by *Shigella*. Mol Microbiol 1995;18:63–75.

[15] Qi J, Wang B, Hao B. Whole genome prokaryote phylogeny without sequence alignment: a K-string composition vector approach. J Mol Evol 2004;58:1–11.

[16] Hao B, Qi J. Prokaryote phylogeny without sequence alignment: from avoidance signature to composition distance. J Bioinform Comput Biol 2004;2:1–19.

[17] Qi J, Luo H, Hao B. CVTree: a phylogenetic tree reconstruction tool based on whole genomes. Nucleic Acids Res 2004;32:W45–7.

[18] Xu Z, Hao B. CVTree update: a newly designed phylogenetic study platform using composition vectors and whole genomes. Nucleic Acids Res 2009;37:W174–8.

[19] Li Q, Xu Z, Hao B. Composition vector approach to whole-genome-based prokaryote phylogeny: success and foundations. J Biotechnol 2010;149:115–9.

[20] Hao B. CVTrees support the Bergey's systematics and provide high resolution at species level and below. Bull BISMiS 2011;2:189–96.

[21] Jeong H, Barbe Y, Lee CH, Vallenet D, Yu DS, Choi SH, et al. Genome sequence of *Escherichia coli* B strains REL6060 and BL21(DE3). J Mol Biol 2009;394:644–52.

[22] Archer CT, Kim JF, Jeong H, Park JH, Vickers CE, Lee SY, et al. The genome sequence of *Escherichia coli* W (ATCC 9637): comparative genome analysis and an improved genome-scale reconstruction of *E. coli*. BMC Genomics 2011;12:9.

[23] Turner PC, Yomano LP, Jarbe LR, York SW, Baggett CL, Moritz BE, et al. Optical mapping and sequencing of the *Escherichia coli* KO11 genome reveal exclusive chromosomal rearrangement, and multiple tandem copies of the *Zymomonas*

mobilis pdc and adhB genes. J Ind Microbiol Biotechnol 2012;39:629–39.

[24] Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, et al. Organized genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. PLoS Genet 2009;5:e1000344.

[25] Oshima K, Toh H, Ogura Y, Sasamoto H, Morita H, Park SH, et al. Complete genome sequence of the wild-type commensal *Escherichia coli* strain SE11 isolated from a healthy adult. DNA Res 2008;15:375–86.

[26] Rasko DA, Rosovitz MJ, Myers GSA, Mongodin EF, Fricke WF, Gajer P, et al. The pangenome structure of *Escherichia coli*: comparative genome analysis of *E. coli* commensal and pathogenic isolates. J Bacteriol 2008;190:6881–93.

[27] Perna NT, Plunkett G III, Burland V, Mau B, Glasner JD, Rose DJ, et al. Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. Nature 2001;409:529–33 [Erratum 410:240].

[28] Hayashi T, Makino K, Ohnishi M, Kurokawa K, Ishii K, Yokoyama K, et al. Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genome comparison with a laboratory strain K-12. DNA Res 2001;8:11–22.

[29] Kulasekara BR, Jacobs M, Zhou Y, Wu Z, Sims E, Saenphimmachak C, et al. Analysis of the genome of the *Escherichia coli* O157:H7 2006 spinach-associated outbreak isolates candidate genes that may enhance virulence. Infect Immun 2009;77:3713–21.

[30] Eppinger M, Mammel MK, Leclerc JE, Ravel J, Cebula TA. Genome anatomy of *Escherichia coli* O157:H7 out breaks. Proc Natl Acad Sci U S A 2001;108:20142–7.

[31] Kyle JL, Cummings CA, Parker CT, Quinones B, Vatta P, Newton E, et al. *Escherichia coli* serotype O55:H7 diversity supports the parallel acquisition of bacteriophage at Shiga toxin phase insertion sites during evolution of the O157:H7 lineage. J Bacteriol 2012;194:1885–96.

[32] Brenner DJ, Staley JT, Krieg NK. Classification of prokaryotic organisms and the concept of bacterial speciation. In: Brenner DJ, Krieg NR, Staley JT, editors. Bergey's manual of systematic bacteriology. 2nd ed., vol. 2. The *Proteobacteria*, Part A. New York: Springer-Verlag. p. 27–32.

[33] Michaescu R, Levy D, Pachter L. Why neighbor-joining works. Algorithmica 2009;54:1–24.

[34] Zuo G, Xu Z, Yu H, Hao B. Jackknife and bootstrap tests of the composition vector trees. Genomics Proteomics Bioinformatics 2010;8:262–7.